Slide 1

IJCAI 99, SVM Workshop, Aug 2,1999

The Analysis of Decomposition Methods for Support Vector Machines

Chih-Jen Lin

Department of Computer Science and
Information Engineering
National Taiwan University

Joint work with Chih-Chung Chang and Chih-Wei Hsu

Slide 2

Outline

- Decomposition methods for SVM
- The convergence proof
- Conclusions

Slide 3

Support Vector Machines

- Training vectors : $v_i, i = 1, \ldots, n$ (length k), a primal dual formulation of SVM:

$$\min \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \eta_i$$
$$a_i ((\phi(v_i)^T w) + b) \geq 1 - \eta_i, i = 1, \ldots, n$$
$$\eta_i \geq 0$$

- $Q_{ij} = a_i a_j \phi(v_i)^T \phi(v_j)$. The Wolf dual:

$$\min \quad \frac{1}{2} x^T Q x - e^T x$$
$$0 \leq x_i \leq C, \; i = 1, \ldots, n$$
$$a^T x = 0,$$

- There are other formulations

Slide 4

Large Dense Quadratic Programming

- $\min \frac{1}{2} x^T Q x - e^T x, a^T x = 0, 0 \leq x \leq C$
- $Q_{ij} = a_i a_j \phi(v_i)^T \phi(v_j)$ in general $\neq 0$, Q : an n by n fully dense matrix
- 30,000 training points: 3GBs RAM.
- Traditional methods: Newton, Quasi Newton cannot be directly applied

Slide 5

Decomposition Methods

- Osuna, Freund, and Girosi [1997a]: separate $x \rightarrow [x_B, x_N]$, fix x_N and work on x_B (like active set method)
- Keep $x_N = 0$, optimize on x_B: obj value =

$$\frac{1}{2}x_B^T Q_{BB} x_B - (e_B + Q_{BN}x_N)^T x_B + \frac{1}{2}x_N^T Q_{NN} x_N - e_N^T x_N$$

- The new sub-problem:

$$\begin{aligned} \min \quad & \frac{1}{2}x_B^T Q_{BB} x_B - (e_B - Q_{BN}x_N)^T x_B \\ & 0 \leq (x_B)_i \leq C, i = 1, \ldots, q \\ & a_B^T x_B = -a_N^T x_N, \end{aligned}$$

Slide 6

- The optimality condition:

$$\begin{gathered} Q_{BB}x_B + Q_{BN}x_N - e_B = \sigma(a_B)_i + \lambda - \mu \\ \lambda_i (x_B)_i = 0, \mu_i (C - (x_B)_i) = 0, i = 1, \ldots, q \\ 0 \leq (x_B)_i \leq C, \lambda_i \geq 0, \mu_i \geq 0, i = 1, \ldots, q \\ a_B^T x_B = -a_N^T x_N. \end{gathered}$$

Note $q \equiv |B|$

- If an $(x_B)_i$ s.t. $0 < (x_B)_i < C$, $\lambda_i = \mu_i = 0 \Rightarrow \sigma$ obtained
- Check x_N if it satisfies the optimality condition
- Move violated components of x_N into x_B
- Q_{BB} is calculated when required: avoid the memory problem
- If $|B| \geq$ # of SVs, no cycles $\Rightarrow$ stop at finite steps
- If #SV large, still not efficient. #SV unknown in priori

Slide 7

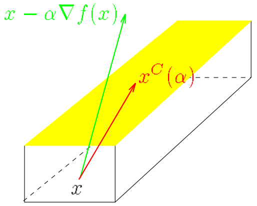
- Osuna, Freund, and Girosi [1997b], fix $q \equiv |B|$ to be small (≤ 100)
  Joachims [1998] ($q = 20$)
  Platt [1998] ($q = 2$)
- Memory problem is solved
- The objective function is decreasing
  More like coordinate search (method of alternating variable)
  Convergence is not clear
- For unconstrained problems, Powell [1973] showed coordinate search may not converge!
- Convergence proof is important as decomposition method presently is one of very few effective methods for large problems

Slide 8

A Traditional Way to Prove Convergence

- Basically, we need sufficient decrease
  For bound-constrained optimization (e.g. Lin and Moré [1998])
  :
  Through Cauchy point

$$x_k^C \equiv P_k(x_k - \alpha_k \nabla f(x_k))$$



Slide 9

Sufficient Decrease

- The Cauchy point satisfies

$$f(x_k^C) \leq f(x_k) + \mu_1 \nabla f(x_k)^T (x_k^C - x_k),$$

and

$$\alpha_k \geq \gamma_1 \text{ or } \alpha_k \geq \gamma_2 \bar{\alpha}_k > 0,$$

where $\bar{\alpha}_k$ satisfies

$$f(x_k(\bar{\alpha}_k)) > f(x_k) + \mu_2 \nabla f(x_k)^T (x_k(\bar{\alpha}_k) - x_k).$$

- x_{k+1} can be any x such that $f(x) \leq f(x_k^C)$
  x_k^C is like an intermediate point

Slide 10

The Convergence

- A sequence $x_1, x_2, \ldots$
  Bounded feasible region $\Rightarrow$ compact $\Rightarrow$ subsequence converges
- For any x in the feasible set (Calamai and Moré [1987])

$$\begin{aligned} & \nabla f(x_k)^T (x_k - x) \\ \leq \; & (\nabla f(x_k)^T (x_k - x_k^C) + \frac{\|x_k^C - x_k\|}{\alpha_k} \max_{x_k + d \in \Omega} \|d\|). \end{aligned}$$

Then

$$\lim_{k \in K_0, k \to \infty} \nabla f(x_k)^T (x_k - x) = \nabla f(x^*)^T (x^* - x) \leq 0.$$

- Decomposition method: $x_k \to x_{k+1}$, only $(x_k)_{B_k}$ is changed

Slide 11

Decomposition Method: Partial Gradients

- Selecting B by solving

$$\begin{aligned} \min \quad & \nabla f(x_k)^T d \\ & x_k + d \in \Omega \\ & |\{d_i \mid d_i \neq 0\}| \leq q \end{aligned}$$

Ω: feasible region

- A modification of Joachims [1998]

$$\begin{aligned} \min \quad & \nabla f(x_k)^T d \\ & a^T d = 0; -1 \leq d \leq 1 \\ & d_i \geq 0, \text{ if } (x_k)_i = 0; d_i \leq 0, \text{ if } (x_k)_i = C \\ & |\{d_i \mid d_i \neq 0\}| = q. \end{aligned}$$

- Note: d is not used as a search direction

Slide 12

- A new definition of Cauchy point

$$x_k^C \equiv P_k(x_k - \alpha_k \tilde{\nabla} f(x_k))$$

where

$$\tilde{\nabla} f(x_k)_i \equiv \begin{cases} \nabla f(x_k) & \text{if } i \in B_k, \\ 0 & \text{otherwise.} \end{cases}$$

We have a more general algorithm

- x_{k+1} is an optimal solution of the sub-problem on B_k
  $f(x_{k+1}) \leq f(x_k^C)$
- $f(x_k^C)$ satisfies similar sufficient decrease conditions

Slide 13

Convergence of Decomposition Methods

- Construct a relation of these two problems

$$\tilde{OP}(x_k) \quad \min \quad \nabla f(x_k)^T d \qquad\qquad OP(x_k) \quad \min \quad \nabla f(x_k)^T d$$
$$x_k + d \in \Omega \qquad\qquad x_k + d \in \Omega$$
$$|\{d_i \mid d_i \neq 0\}| \leq q$$

Then for any x in the feasible set

$$\begin{aligned} & \nabla f(x_k)^T (x_k - x) \leq -OP(x_k) \leq -n^2 \tilde{OP}(x_k) \\ = \; & n^2 \max_{x_k + d \in \Omega_k} -\tilde{\nabla} f(x_k)^T d \\ \leq \; & n^2 (\tilde{\nabla} f(x_k)^T (x_k - x_k^C) + \frac{\|x_k^C - x_k\|}{\alpha_k} \max_{x_k + d \in \Omega} \|d\|). \end{aligned}$$

Slide 14

Conclusions

- Many optimization issues from SVM have not been studied yet
  Both theoretical and numerical aspects
- The interaction between SVM and optimization researchers is important for solving large-scale problems